EUROCONFERENCE

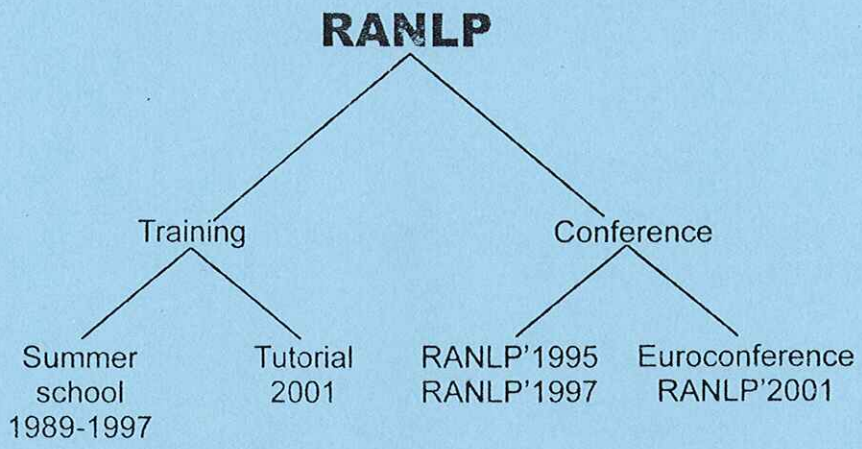
RECENT ADVANCES IN

NATURAL LANGUAGE PROCESSING

Supported by the European Commission, DGXII, Human Potential Programme,
High Level Scientific Conferences, Contract number HPCF-2000-00329

PROCEEDINGS

Edited by
Galia Angelova, Kalina Bontcheva, Ruslan Mitkov, Nicolas Nicolov, Nikolai Nikolov



Tzigov Chark, Bulgaria

5-7 September 2001

TABLE OF CONTENTS

REGULAR AND SHORT PAPERS

Adapting the Generalized Upper Model to Spanish

Mª del Socorro Bernardos Galindo[1] and Guadalupe Aguado de Cea[2]

[1] Laboratorio de Inteligencia Artificial,
[2] Departamento de Lingüística Aplicada a la Ciencia y a la Tecnología,
Facultad de Informática, Universidad Politécnica de Madrid.
Campus de Montegancedo, s/n. 28660 Boadilla del Monte, Spain.
sgalindo@delicias.dia.fi.upm.es
lupe@fi.upm.es

Abstract

The domain information source does not usually contain the linguistic knowledge that a natural language generation (NLG) system requires to be able to produce texts. This paper presents the solution adopted in a project that generates texts in Spanish. We decided to use an ontology which provided that linguistic knowledge. Then we could classify the domain information under the new ontology, so that it could inherit that knowledge. Instead of starting from scratch, we reused an existing ontology, called *Generalized Upper Model* (GUM), that had already been successfully used for text generation in other languages. We studied each GUM category in depth and made suitable modifications, according to some criteria determined at the beginning of the process. These criteria helped to ensure that the ontology was application and domain independent.

1 Introduction

Natural language generation (NLG) systems use several different kinds of information: about the domain, the user, the context, the lexicon, the grammar, etc. This paper is only concerned with aspects related to the domain information.

The domain information source (knowledge base, database, etc.) is not a constituent part of an NLG system as such, but it is a fundamental resource for it, since it provides the system with what it can say or has to 'say'. Traditionally there has been a distinction between the NLG system and the application that uses it. Figure 1 shows a graphical representation of their interaction with the domain information source.

Domain data are usually different from the terms an NLG system requires. This makes it necessary to determine a mechanism that maps the (non-linguistic) domain terms onto the (linguistic) terms that an NLG system can use.

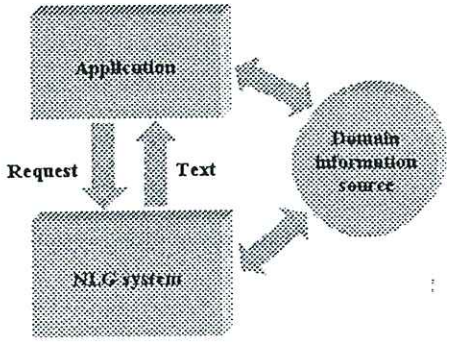


Figure 1: Relationship among the application, the NGL system and the domain information system.

Bateman (1996) collects some of the possible solutions to this problem:

- Making a domain-specific encoding of how the application domain requires its information to appear.
- Setting up mapping tables from the domain categories to the linguistic resources of the generator.
- Doing a progressive replacement of component configurations of the domain representation with other configurations that are nearer (in structure or content) the surface structure to be generated.
- Reducing the possible mappings between domain and NLG-terms to just the relation of logical subsumption, by classifying the domain categories in terms of a hierarchy of general objects and relations that behave systematically with respect to their possible linguistic realizations. Such a hierarchy is provided, for example, by the *Generalized Upper Model (GUM)* (Bateman *et al.*, 1995a).

The case presented here describes the steps followed in an NLG project, hereafter referred to as ONTOGENERATION[1] (Aguado *et al.*, 1998), whose aim was the construction of a system that answered in Spanish to queries about the chemical elements and their properties. In this project we opted for the last solution mentioned above, namely, to use a GUM-like ontology.

[1] Project funded by Universidad Politécnica de Madrid, under the call "Ayudas para la realización de proyectos de investigación y desarrollo, dirigidas a grupos potencialmente competitivos" (Reference A9706).

Next sections give a short overview of GUM, explain the reasons to choose it, and describe the process carried out so as to be able to use it in the NLG system. Due to limits of space, this paper does not deal with the generation process or the implementation aspects.

2 Overview of GUM.

GUM is a linguistic ontology, bound to the semantic of the constituents of a language grammar. Unlike other linguistic ontologies, such as WordNet (Miller, 1995), it does not describe the semantics of words, but the semantics that can be expressed in bigger grammatical units, such as nominal groups, prepositional phrases, etc.

GUM consists of two hierarchies: one of concepts and one of relations. The concept hierarchy represents the basic semantic entities and includes configurations of the processes and the different kinds of objects and qualities. The relation hierarchy represents the participants and the circumstances involved in the processes, and the logical combinations between them. Figure 3 shows the first levels of these hierarchies. These taxonomies have their origin in Halliday's (1985) functional grammar, but can be applied to any theory.

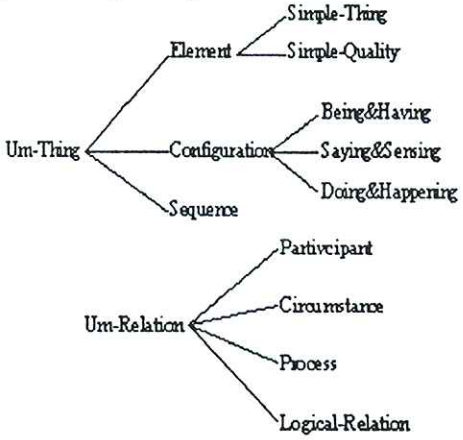


Figure 3: First levels of GUM hierarchies

The categories considered in the hierarchies can belong to one of the following types: dimension (to represent different points of view about an entity), partition (to represent mutually exclusive alternatives that cover all the possible specializations of an entity), disjunction (to represent mutually exclusive alternatives that do not cover all the possible specializations of an entity), and simple specialization (to represent distinctions about which we cannot state anything about their disjunctive properties).

3 Reasons to choose GUM

There were two main reasons to choose GUM as a base to develop the interface between the NLG system of ONTOGENERATION and its domain information source.

First of all, previous work using GUM had shown that it could supply a solid basis for providing natural language generation capabilities when domain organization was isolated from the details of its linguistic realization. Thus, using GUM as an interface level ensured that:

- We did not have to import linguistically-motivated distinctions into our domain ontology (an ontology called *Chemicals* (Fernández, 1996)) in order to support natural language generation, because when we classify domain concepts according to the GUM conceptualization, domain terms inherited the possibilities for their linguistic expression from the ones in GUM. This way, we did not interfere with the domain-internal organization or violate the desirable modularity of a complete system (Lang (1991) critizes such a violation in the LILOG project).
- Making an interface with our grammatical realizer[2] (a tool called KPML (Bateman, 1997)) was simplified, because much of the information specific to language processing was bound to the relationship between GUM and the linguistic resources and, therefore, it was not necessary to include it in the input specifications required by the realizer.
- The need for domain-specific linguistic processing rules was reduced, because GUM provided that knowledge via a domain-independent general, reusable conceptual organization that could be used to classify all the domain specific knowledge.

The second reason was the possibility of reusing GUM successfully, as had already been shown in previous work for several languages. GUM provided a fixed anchor that was sufficiently general as to require only minor variations across languages. It was not necessary to adopt an interlingual position, although it was still possible to minimize the language-specific idiosyncratic aspects of the semantic descriptions. The existence of a suitable documentation ((Bateman *et al.*, 1994), (Bateman *et al.*, 1995a), (Bateman *et al.*, 1995b), (Henschel, 1993), (Magnini, 1994)) made this reuse an easier process.

Since, at the beginning of the project, GUM only supported generation in English, German and Italian, we had to study and then adapt it to Spanish, by creating a new ontology that we called GUME.

Next sections try to reflect the main features of that adaptation.

4 Design criteria to develop GUME

The criteria adopted to construct GUME were based on the guidelines followed during the development of GUM, from the first version, created for the Penman system (Bateman *et al.*, 1990), to the last one; we also added some others, the last three ones, to improve the structure and understanding of the model. These criteria were as follows:

[2] A grammatical realizer translates a phrase specification into a grammatically correct sentence. The complexity of the realizer's task will depend on the level of sophistication of the specifications (templates, abstract syntactic structures, etc.).

- The concepts and relations of the ontology would be necessary to represent distinctions in their lexico-grammatical realization and reflect differences in their experiential meaning[3].
- The types of hierarchical divisions would be the same as in GUM: dimensions, partitions, disjunctions and simple specializations.
- Configurations with a different number of participants and circumstances would have different conceptual representations.
- The syntactic variation that involved the order of the arguments would only be relevant when the pattern difference caused a change of meaning.
- An expression would be allowed to be generated from different categories, according to the semantic perspective taken.
- The style would be uniform. That would involved, for instance, the modification of the divisions of some categories, avoiding the combination of dimensions with simple specializations.
- The changes would be minimized. There were two reasons for this: GUM (or a previous version) was already being successfully used, and if there were only a few differences between GUME and GUM, their integration would be easier.
- New categories would receive an English name in order to achieve a uniform terminology.

When some criteria were in conflict with others, each case was studied separately and the best option was chosen.

5 Method to develop GUME

As we parted from an existing ontology, we followed a construction method that allowed us to take the most of the work already done (in this case, GUM).

The process was the following: for each category of GUM, we studied its description thoroughly. Then we looked for a set of Spanish linguistic behaviours equivalent to those represented by the category and compared these behaviours with the explanations given in GUM. If we observed discrepancies, we proposed and made the corresponding extensions[4], that would have to meet the design criteria. We also studied organizational alternatives and if there was one better, we made the appropriate modifications.

In order to build GUME, we used another resource, besides GUM: a corpus with examples of the texts that the system would generate (Bernardos & Aguado, 2001). The use of the corpus could serve two purposes:

- To find out if a new modification was necessary. For example, in the case a text did not fit in the ontology.
- To help us validate the ontology, by confirming that, at least in the chemical domain, GUME was correct. Since we only used texts from a particular domain and there was not a linguistic theory for Spanish, equivalent to the one used for GUM, we could not guarantee its validity for all domains. The systematic criteria and method adopted let us assume that changes, if needed, would be minimum.

At this stage, the procedure followed was: for each text in the set, we looked for GUM categories where each of its components could be classified. If some part of the text could not be classified within GUM, or could only be classified in a very abstract category, and it was a general linguistic expression in Spanish -i.e. it did not belong only to the chemical domain, we created the category required to represent the meaning of that part.

6 Results

In order to illustrate the previous section, we present first a numerical summarization of the result obtained in the adaptation process, followed by some examples of different categories.

- We identified 185 completely valid GUM categories (119 concepts and 66 relations).
- We identified 6 not valid GUM categories (3 concepts and 3 relations).
- We identified 20 GUM categories that needed modifications (12 concepts and 8 relations).
- We identified and implemented 15 new categories (11 concepts and 4 relations).
- The whole GUME ontology consists of 142 concepts and 78 relations.

As we expected, GUM provided a conceptualization that was general enough so as not to require many changes for Spanish.

6.1 Example of category creation.

In GUM there was a concept, called, *Generalized-Positioning*, used to represent situations where an entity is located in space or time. Some sentences reflected in this category are:

English: John is at the station.
The meeting is at 3.30
The concert is near here.
Spanish: John está en la estación.
La reunión es a las 3.30.
El concierto es cerca de aquí.

While in English only one verb ("to be") is used, in Spanish there are two ("ser" and "estar"). We use one or the other depending on the entity being located. If the entity is an event, we used "ser" and if it is an object, we use "estar" (Seco, 1990).

This led us to create two specializations of this category: *Generalized-Positioning-Event* and

[3] According to Halliday, a sentence has contributions from different metafunctions, each one giving rise to a set of associated constraints. The experiential metafunction is one of them.

[4] These differences were usually related to the division of the subhierarchy and its associated linguistic constraints.

Generalized-Positioning-Object, which have proved to reflect this specificity of Spanish.

6.2 Example of category elimination.

GUM has a category, called *Name-of*, that reflects the relation between a name and the entity that bears or "has" it. After its analysis, we did not observe any differences in Spanish between the possession of an object (concept *Ownership*) and the possession of a name. Furthermore, there is a category, called *Name-Relation*, that represents expressions of the kind: *X is called (X se llama Y)*, that also correspond to this category. We decided to eliminate this category[5].

It is worth pointing out that it also seems to be unnecessary for English[6].

6.3 Example of category modification.

The category *Generalized-Role-Relation* reflects the generalized perspective about an entity that plays a role with respect to another. In GUM, the following expressions are represented by this category:

English: <domain> has <range> as <role>
<domain> 's <role> is <range>
Spanish: <dominio> tiene <ámbito> de <papel>
el <papel> del <dominio> es <ámbito>

For instance:

English: John has Betty as a secretary.
Spanish: John tiene a Betty de secretaria.

However, sentences such as *John 's secretary is Betty* (*La secretaria de John es Betty*) are also said to be generated from the concept called *Identity*. This sentence maps to the pattern *<domain> 's <role> is <range>*. After some discussion, we concluded that it was an identity, since *John's secretary* and *Betty* are the same person, so we eliminated the second option from the category *Generalized-Role Relation* in GUME[7].

6.4 Example of the same category.

Both GUM and GUME use a category, called *Existence*, to represent configurations concerning only one entity. This category reflects that there is something or that something exists, for instance:

English: There is a block.
Spanish: Hay un libro.

The way of expressing the existence concept is very different in both languages. This does not involve that two different categories are needed, but that their linguistic constraints would be different in their corresponding grammar.

7 Link between the domain information source and GUM(E).

As pointed out in the introduction, the link between the domain information and GUM (in this case, GUME) can be achieved by classifying that knowledge in terms of the general semantic categories provided by GUM(E). Once a domain category is subsumed under a GUM(E) category, the system can make inferences about how that domain category must be expressed, since it inherits linguistic constraints from the corresponding GUM(E) category.

Some researchers disagree with this approach, i.e. with the theory that the domain knowledge has to be subsumed under GUM(E) categories (see (Stede & Grote, 1995) and (Stede, 1996)). Their reason is that mapping between GUM and domain knowledge can involve a complex restructuring of the domain source, making the subsumption unfeasible.

In ONTOGENERATION, the taxonomic structure of *Chemicals* avoided the problem just mentioned. The subsumption process would not present much trouble: we would just have to link the root of the chemical ontology under the concept *Object*.

8 Conclusions and future work

In this paper we have presented the solution adopted in an NLG project in Spanish to construct an interface between the domain information source and the generation system. We decided to reuse a linguistic ontology, called GUM, that had been previously used in other NLG projects and for different languages. Thus, our project shows its usefulness for Spanish.

Besides, the criteria and procedures followed to adapt GUM to Spanish are application independent and can, therefore, be applied to other languages.

We would like to point out that the resulting ontology is not very big (see section 6), especially compared to other linguistic ontologies for Spanish, such as Eurowordnet (Díez *et al.*, 1997). This is due to the fact that it does not contain information about the word semantics, but about bigger grammatical units. For that reason, in ONTOGENERATION, we had to introduce data about general terms, such as articles, verbs, etc., as well as a lexicon with chemical domain terms.

However, the use of GUM and, therefore, of GUME, present other advantages. First it helps to keep the modularity of the system. On the one hand, we can separate the general domain knowledge from the linguistic details and, on the other hand, we can simplify the communication between the modules, as part of the information needed to generate the texts is bound to the ontology and it does not need be passed from one module to another. In relation to this, we can state that the complexity of the generation tasks is reduced, especially

---

[5] This is not in conflict with criterion 5. We did not eliminate *Name-of* because there was another category that served the same purpose, but because we thought that it should not be an specialization of *Generalized-Role-Relation*. The existence of *Name-Relation* made unnecessary the creation of a new category.

[6] This fact was discussed with Dr. Bateman, major creator of GUM, so that it could be taken into account in future versions of the ontology.

[7] This change was also presented to Dr. Bateman, in order to be considered in future versions.

the syntactic selections. Finally, GUM(E) can also be very useful for multilingual applications with multilingual tools such as KPML.

We would also like to indicate that using examples of domain texts during the process of adaptation of GUM to Spanish does not make GUME domain dependent, since those texts are only used as an aid for the development of the ontology, not as a guideline. The procedure was intended to ensure that domain specific data were not included. The use of this model in the chemical domain has proved its validity in that domain, but we are trying to apply it in other projects to find out if the eliminated, modified and inclùded categories are really correct, and if new changes are needed.

Regardless the evaluation provided by the use of the ontology, there are several categories in GUM(E) that deserve a deeper analysis:

- The configuration called *Doing&Happening* reflects a wide range of sentences and has only a few specializations. Very different semantic expressions, such as *La casa se derrumbó (the house fell down), Ha muerto (he has died)* or *Surgió de repente (It appeared suddenly)*, belong to the same category. It would be useful to study the convenience of adding new categories to this configuration.
- The concept *Sequence* is still under examination in the development of GUM, so it is probably the part of the ontology that will evolve above the rest in the future.

Finally, one of the most interesting tasks is the integration of GUME and GUM, making the resulting model useful for multilingual NLG systems that include English, German, Italian and Spanish.

References

(Aguado *et al.*, 1998) G. Aguado, A. Bañón, J. Bateman, S. Bernardos, M. Fernández, A. Gómez, E. Nieto, A. Olalla, R. Plaza and A. Sánchez. ONTOGENERATION: Reusing Domain and Linguistic Ontologies for Spanish text generation. *Workshop on Applications of Ontologies and Problem Solving Methods, ECAI'98*, Brighton (UK), 1998.

(Bernardos & Aguado, 2001) S. Bernardos and G. Aguado. A New Approach in Building a Corpus for NLG. *Computational Linguistics and Intelligent Text Processing*. Gelbukh (ed.). Springer-Verlag, pp. 216-225, 2001.

(Bateman *et al.*, 1990) J. A. Bateman, R. T. Kasper, J. D. Moore and R. A. Whitney. A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model. *Technical Report*, USC/ISI, Marina del Rey, (USA), 1990.

(Bateman *et al.*, 1994) J. A. Bateman, B. Magnini and F. Rinaldi. The Generalized {Italian, German, English} Upper Model. *Proceedings of the ECAI'94*, 1994.

(Bateman *et al.*, 1995a) J. A. Bateman, R. Henschel and F. Rinaldi. Generalized Upper Model 2.0: documentation, *Technical Report, GMD/IPSI*, Darmstadt (Germany), 1995.

(Bateman *et al.*, 1995b) J. A. Bateman, B. Magnini and G. Fabris. The Generalized Upper Model Knowledge Base: Organization and Use. *Towards Very Large Knowledge Bases*, 60-72. IOS Press, 1995.

(Bateman, 1996) J. A. Bateman, Automated Discourse Generation. *Encyclopedia of Library and Information Science*, 1996.

(Bateman, 1997) J.A. Bateman. Enabling technology for multilingual natural language generation: the KPML development environment. *Natural Language Engineering*, 1: 1-42. Cambridge University Press, Cambridge (United Kingdom), 1997.

(Díez et al., 1997) P.Díez, W. Peter and P. Vossen. The Multilingual design of EuroWordNet. *Proceedings of ACL/EACL-97. Workshop on Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*. Madrid (Spain), 1997.

(Fernández, 1996) Fernández, M.. *Chemicals: Una Ontologia de Elementos Químicos*, Trabajo Fin de Carrera, Facultad de Informática, Universidad Politécnica de Madrid (Spain), 1996.

(Halliday, 1985) M. A. K. Halliday. *An Introduction to Functional Grammar*. Edward Arnold, London (UK), 1985.

(Henschel, 1993) R. Henschel. Merging the English and the German Upper Model. *Technical Report*. GMD/IPSI, Darmstadt (Germany), 1993.

(Lang, 1991) E. Lang. The ontology Lilog from a linguistic point of view. Text understanding in Lilog: integrating computational linguistics and artificial intelligence. *Final report on the IBM Germany Lilog-Project*, Springer-Verlag, pp. 464-481, 1991.

(Magnini, 1994) B Magnini. Specification of Upper Model, *Technical Report Project 062-09 GIST*. IRST, 1994.

(Miller, 1995) G. A. Miller. WordNet: a lexical database for English.. *Communications of the ACM* 38 (11), pp. 39 – 41, 1995.

(Seco, 1990) Seco, R. *Manual de Gramática Española*, Aguilar, Madrid (Spain), 1990, 11th edition.

(Stede & Grote, 1995) M. Stede and B. Grote. The lexicon: Bridge between language-neutral and language-specific representations. *Working notes of the IJCAI workshop on multilingual text generation*, Montreal (Canada), 1995.

(Stede, 1996) M. Stede. *Lexical Semantics and Knowledge Representation in Multilingual Sentence Generation*. PhD. Thesis, University of Toronto (Canada), 1996.